Dataset compilation

Area of focus

As can be gathered from the Argumentative Task Design document, the proposed dialogue system would be a complex tool consisting of several subsystems to control the numerous subtasks. It would also requirelarge amounts of data to be trained to analyze students' utterances, as well as to retrieveor generate counterarguments and point to the most relevant information sources. Given the complexity of the system and the small scale of our project, we have decided to startour tests focusing on subtasks 3 and 4 (ST3 and ST4), where the student selects the paragraphs with the thesis and an argument, and where the student summarizes the author's thesis and main argument. Our tests involve compiling a dataset with, first, annotated texts where the student could be asked to select the paragraphs with the thesis and an argument (ST3),and, secondly, possible answers that the students could give when prompted to summarize the thesis and argument (ST4). The tests also involve carrying out an experiment using the dataset to determine whether a selected tool could help the system analyze students' answers to ST4 in order to provide adequate feedback.

It might seem more logical to start from the beginning, ST1, but since that part is concerned with guiding the student to complete the task, it might be better to tackle this aspect after having at least a first version of the rest of the system. It might then seem that ST2 is the second most logical place to start. However, experiments regarding this part would require a voice dataset, which would require access to a sufficiently large poolof subjects – especially given the idiosyncrasy of speech recognition with young subjects (Yu et al.). For this same reason, we are performing our experiments using only text inputand output, without adding voice. The next subtasks, ST3 and ST4, can be studied more easily with artificial [1]written data; we believe that this imperfect data might suffice for developing a first sketch of these system components. Further experiments with real datamay prove or disprove the usefulness of these initial efforts – at any rate, we hope that wecan provide the structure for the ulterior collection of real data.

Dataset justification

Dialogue systems can generally be described as functioning based on rules or corpus data(Jurafsky and Martin, 2019). Some rule-based systems can function adequately, such as the popular ELIZA chatbot (Heller, 2016; Jurafsky and Martin, 2019), but without corpus data a chatbot's capabilities are very limited[18]. Even if the architecture relies solely on hand-written rules, the people writing the rules will need some data to use as reference. Thus, even for a small project like ours, carrying out preliminary tests for a component of a simplified system, data is essential. We need data in the form of texts that the students would analyze (ideally a dataset of texts adequate for students, with their argumentative structure annotated in a simple but informative manner), and as

[1] We refer to this data as artificial because, though it was written by a person, it did not come from the target users of the system nor was it written under the target conditions of completing an educational task

sample answers that thestudents would give when prompted to analyze the texts (ideally brief yet complete thesesdealing with the same precise topics as the texts from the text dataset).

As mentioned when discussing the task design, we selected the SAT test as a good starting point forour task – SAT tests provide us with appropriate texts that can serve as basis for thetask that the system is to perform with the students. However, the SAT Essay requires students to write a complete essay analyzing all the resources that the authors uses in their argumentation, but never discussing their own opinion (CollegeBoard, 2015). This means that SAT Essay answers would certainly not be useful for ST5, where students are askedto develop their own argumentation. The SAT Essay answers are also not useful to train the system to evaluate student's answers to ST3 and ST4 for two reasons:

Structural differences: In the SAT Essay test, the student answers by writing their analysis as an essay. The purpose of our task is not merely evaluative, but primarily formative, which means that the task is broken down into simpler tasks, providing the student with the necessary scaffolding.

Availability: Even though test prompts are readily available, the same is not true for test answers, whether scored or unscored.

We looked for suitable data in several datasets, but rejected them. We considered datasets by IBM, the args.me corpus, datasets from ArguAna and the Ubiqutous Knowledge Processing Lab. One reason why we rejected them was the complexity of the annotations, not suitable for comparisons with analyses to be made by high-school students. Another reason was that they contained very short texts or texts of inadequate quality to serve as educational models or as basis for an analysis. Some datasets were also rejected because they contained isolated arguments or claims, not complete argumentative texts - these could however be useful for ST5, where the system is to analyze the students' own arguments and respond with suitable counterarguments. Despite the limitations of using only the SAT texts, it was considered the best option, if only as basis for the task.

Having selected the text data that the students would analyze, the next challenge wasobtaining sample answers that they could give. As mentioned above, the datasets that we found were not suitable. A Wizard-of-Oz experiment would be the ideal means of obtaining a dataset to train the system (Thies et al., 2017). In this ideal setting, a large and diverse sample of target students would complete the task on a computer, believing they are talking to a dialogue system, but it would be a person, trained for the task, who would be talking to them. This would return data that could be considered real data, as it would have been produced by target users in a context made to simulate the target settingas closely as possible. However, such an experiment would be costly. Large, high-quality public datasets would have been another good alternative. As we mentioned, we looked at argument-mining datasets to see if they could be used to train our system. However, given the specificity of our task, available datasets did not seem useful: some were only annotated for topic but not stance, the ones annotated for quality followed criteria not aligned with ours, the topics covered did not match the SAT tests, and none were obtained in the context of high-school students completing a task. For these reasons, it seemed more feasible to perform our tests with an artificial dataset created

by ourselves. Writing our own answers would allow us to control the variables that we wanted the system to analyze, both in terms of type and sample size. Naturally, artificial answers cannot be expected to perfectly resemble real data, but they can serve as a starting point to developthe basis of a system which can then be tested on real users to obtain real data that can be used to further develop the system (Jurafsky and Martin, 2019). Additionally, despite the limitations on the answers part of the dataset, the text data is from the source we believe to be most appropriate, as justified when discussing the task design. Thus, we provide a solid base that might be of sufficient quality for future data gathering through Wizard-of-Oz studies.

Argumentation model

Naturally, the analysis of argumentative texts requires a model to guide it. One of the most influential models is probably Toulmin's (Toulmin, 2003, in Andrews, 2005). Andrewsplaces it around the middle of a spectrum that ranges between formal logic and rhetoric. This middle point is believed to be appropriate for education, where argumentation occurs in different disciplines and contexts (ibid). Argumentation is described as a process linkedto "transformation, clarifying and changing ideas, personal growth, identity formation, and other dynamic aspects of learning", not bound by the conventions of a single textual genre (ibid, p. 110); especially this last point makes it necessary to adopt a flexible model,far from the rigidity of formal logic, but with enough structure to be used by a computer system.

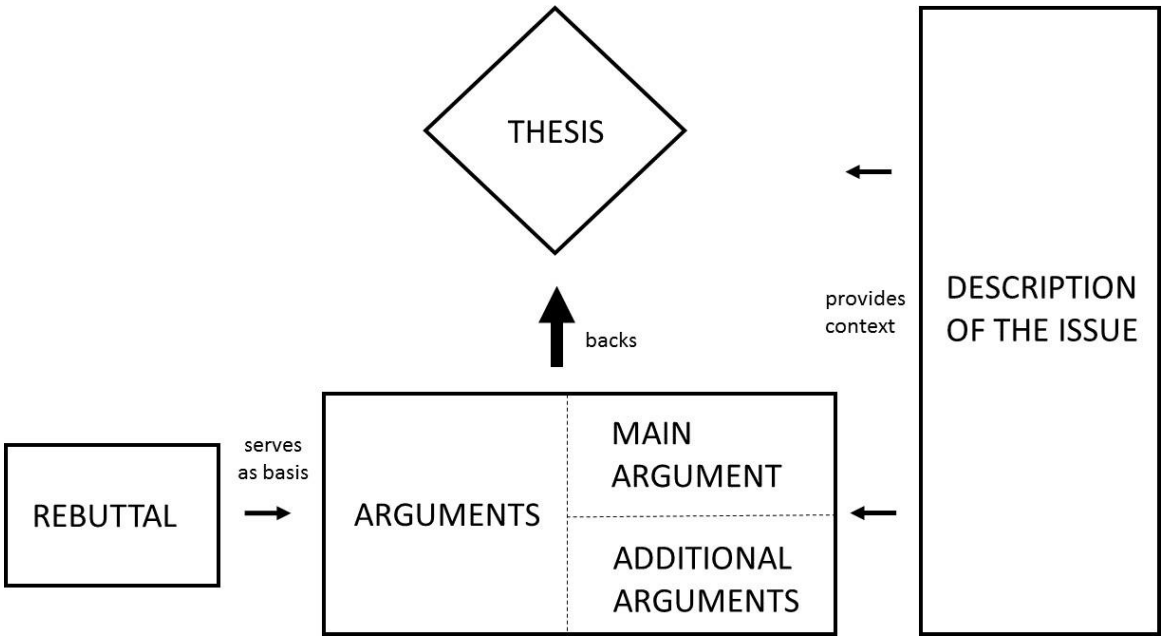

Figure 1: Argumentation model used to annotate our SAT texts

In Toulmin's model, the main elements of an argument are claims (defending a position), grounds (evidence to justify defending that position), warrants (linking the evidence to the claims), backings (justifying the warrants) and possibly qualifiers (modifying the link between grounds and claims) and rebuttals (challenging the relationship between

grounds and claims) (Andrews, 2005; Stab and Gurevych, 2014).

It seems necessary to arrive at a highly simplified model, due to the level of the target audience of the dialogue system (high-school students), the limited capabilities of a simpledialogue system and the difficulty of annotating arguments coherently (Gurevych, 2021). Stab and Gurevych (2014) provide a useful example of such a model in their guidelinesfor the Argument Annotated Essays Corpus (AAEC) (ibid); we take their model as an important reference because it is intended for use in Computational Linguistics. They later describe a more complex structure, which is justifiable given the target users of their dataset (computational linguists). However, their description of the elements of an argument begins with only claims and premises to back the claims. They later establisha hierarchy, distinguishing between claims and major claims. In our annotation scheme, we thus establish as the main elements the thesis (the major claim) and the arguments that support it (the premises). This structure may suffice to classify all the SAT textswe encounter, though some texts and text chunks require departing from this simplest of models and adding some more elements, while attempting to not complicate the model beyond what we consider essential. Figure 1 shows the final form of our selected model; below we explain the modifications we made to the initial simple model to arrive at the model shown in the figure.

One element that we have deemed necessary to add is the one we have labeled "description of the issue". Argumentation models, at least when intended for argument analysis (production lies beyond the scope of our experiments), are models for "distilling the salientfrom the residual" (i.e. separating the elements that contribute to an argument from other text parts that do not serve that purpose) (Andrews, 2005, p. 114). In order to provide thebest feedback possible, we think that we also need to label these "residual" elements (at least as such, perhaps without analyzing them any further), so that if a student focusedon the residual rather than on the salient we may inform them of the nature of their error.Aiming for the utmost simplicity, we have labeled the "residual" elements (i.e. not directlyinvolved in the argument) as "description of the issue", since we have found that they generally serve the purpose of setting the context for the reader; it may be argued that they serve additional purposes, such as appealing to the reader's emotions with personal exam-ples, but we believe that the "description of the issue" label encompasses most "residual"content in the simplest, broadest manner, including the examples. Still, we have reservedan "other" label for "residual" elements that do not fit the previous label (e.g. we have found a text where one paragraph tells the reader how to stay informed and engaged withthe issue).

As we have described, the model by Stab and Gurevych (2014) includes hierarchical elements. While we tried to avoid complicating our model this way, it seemed justified toestablish a hierarchy in some texts. We have observed some texts where the author seems to emphasize one argument over others; in such cases, we believe that a true understandingof the text on the part of the student would require seeing this distinction. We have thus decided to distinguish between main arguments and additional arguments in our annotation, but only where the author gives more importance to a particular argument, by mentioning it alongside the thesis and devoting more paragraphs to it than to other arguments. Another reason to justify this decision is that the "additional arguments" might not even be considered arguments in a more complex model, such as Toulmin's fullmodel (Andrews, 2005; Stab and Gurevych, 2014), where they might be warrants linking the grounds to the claims. For instance, the May 2018 SAT text claims that research companies should make their data accessible to others; the main reason given for this is that medical research requires having all the

data. One of the additional arguments given is that many of the studies that are made accessible do not include their data; in our simple model this has been labeled as "additional argument", but in a more complex model this could be seen as a warrant linking the need to publish data (claim) with the need fordata in medical research (premise). Another example is the October 2018 SAT text about dwindling bat populations, which mentions both at the beginning and the end that bats are important to us because they eat insects. Thus, the argument about the value of bats to humans can be considered the main argument, while other arguments are mentioned once at different points in the text, but not emphasized in the introduction or conclusion (e.g. the argument about it not being costly for energy companies to make adjustments that would prevent bat deaths or the argument that, while curing diseases that affect batsis difficult, helping bats by preventing other causes of death is within our hands).

It must be noted that this distinction between main and additional arguments is based solely on the importance given to them by the author, not in their effectiveness. In ST3 and ST4, the student is not meant to evaluate the effectiveness of the arguments, only toidentify them. This may reduce the learning gains from these subtasks, but this may be compensated by the following subtask (ST5), where the student does need to use critical thinking to evaluate their stance on an issue, present and support arguments and respondto counterarguments. At any rate, it must be born in mind that this is meant to be a scaffolded activity; only after students have developed their skills further can the scaffold be removed, and more demanding tasks be presented to the students.

One last element that we needed to add to our simple model is the "rebuttal" label; this is one of the elements of Toulmin's full model (Andrews, 2005; Stab and Gurevych, 2014). In the earlier, simpler version of our model, we could consider any rebuttals as part of the argument that uses them as basis to challenge a view opposed to the author's. For example, the text from the December 2016 SAT mentions how some opposed giving workers unpaid leave for family reasons claiming that it would be bad for businesses; the author then quotes statistics on how this did not harm businesses, as a way to support their argument in favor of a new law making that unpaid leave paid. A complex argumentation model could separate the argument where some people believed unpaid leave to be bad for businesses from the argument where the author claims this turned out to be false; in our simplified model, we opted for considering both elements as part of the same idea that giving workers certain rights may not harm businesses in the end. Nonetheless, given thatwe established the paragraph as our working unit (as we describe in the next section), it became necessary to add the "rebuttal" label for cases where the rebuttal and the argument ensuing from it are in separate paragraphs. For example, this happens in theMay 2019 text (included as C). There, the author speaks in favor of a law requiring firmsto disclose possible gender pay gaps. The author chooses to use a structure where they first present any counterarguments (e.g. paragraph seven covers most of them, saying that some believe the measure to be costly, complex and ineffective), to then contrast them with their arguments in favor of the measure in the following paragraphs (e.g. paragraph eight gives an example of a company where the measure worked well). This label could alsoprove useful in the unlikely but theoretically possible case where an author mentions some views opposed to the one they are defending and does not challenge these rebuttals (e.g. though an unchallenged rebuttal might weaken an argument, readers might appreciate an author acknowledging the limitations of their claims, especially when the author explicitly defends the "lesser evil", i.e. the less limited claim). Whichever the case, it must be noted that this label is only added for the system to be able to give adequate feedback: for example, if the

student points to a paragraph containing only a rebuttal thinking that it isa supporting argument, this label can help the system explain to the student the nature of their error. That would be the only use for this label, as it is not part of what the student would need to analyze in the task: the task is concerned with the student identifying thekey ideas of the text (thesis + arguments), not analyzing its entire structure.

Annotation unit

We have chosen the paragraph as our unit for annotation. We justify this decision withfour reasons: avoiding conflicting views about how the texts are structured, adapting tothe students' level, saving time on the annotation process, and respecting the author's choices. These reasons are further explained below.

Firstly, opting for larger units reduces disagreements about where elements of an argument begin and end; what matters in our system is simply where they are contained, nottheir precise beginning and end. The mere fact that Stab and Gurevych (2014) had to gointo great detail in their annotation guidelines to explain to experts how the argumentative elements were identified suggests that this is highly subjective and complex. When interacting with a simple dialogue system, students cannot debate the correctness of their answer as they could with a human partner; therefore, and to spare students the frustration of being corrected when they are technically right (Kulatska, 2019), the system's data needs to be simplified to increase the likelihood of feedback being appropriate.

The annotations would be used by the system as a gold standard to compare students' answers with. Annotating units smaller than a paragraph would require a good knowledge of syntax - this level could be expected from expert annotators, but perhaps not from high-school students who may never have taken a Linguistics class.

Using a large unit like the paragraph reduces the time needed to annotate the texts and thus enables us to produce a larger dataset, which may be more useful than a smallerone. It must also be noted that the SAT Essay texts are longer than the texts annotated in most argument mining datasets. Moreover, paragraphs are normally separated by para-graph breaks, a visually noticeable separation that could make it easier for students and annotators to distinguish between units in the text.

The author of the text divides it into the paragraphs they see fit; therefore, this divisioncan be expected to closely match the author's idea of which elements form the text.

Assigning labels to paragraphs

Ideally, for the data to be simple and thus ease the annotation task, the design of the system and the ulterior educational task, each paragraph of the text should be annotated with just one label. For example, if a paragraph includes a counterargument to the author's thesis and immediately debunks it, adding a "rebuttal" label to the "argument" label would serve no purpose: the student would only be asked to spot the arguments, so thatis all the annotation that the paragraph would need; other labels would only be necessaryin paragraphs performing functions different from what the student would have to spot,so that the system could better explain to the student why they pointed to the wrongparagraph. However, it is not always possible to assign only one label to a paragraph. Inthe dataset annotation process, the following cases have been found where a paragraphmay perform more than one function:

A paragraph contains both a counterargument and an argument that addresses that counterargument: it is easier to consider that counterargument as an element to back the argument, rather than another isolated argumentative function. Thus,despite the paragraph performing more than one function, we assign it only the "argument" label. As we explained when discussing the argumentation model, this is what happens in the December 2016 text, where, in the same paragraph, the author mentions how some people believed unpaid leave to be bad for business, and then gives statistics that prove that to be false. In some cases,however, the rebuttal and the author's argument addressing that rebuttal are written in separate paragraphs; then the paragraph with the rebuttal is given the "rebuttal" label. This is the case of paragraph seven in the May 2019 text, as we detailed when discussing the annotation unit.

A paragraph includes more than one distinct argument: when those arguments are very similar, we prefer to group them to simplify the task at all levels (annotation, development of the system, completion by the student). An example can be found in the May 2019 text, where the author talks in favor of companies disclosing gender gaps. In paragraph nine, they mention a company where this measurewas implemented, and this helped them see why there was a gender gap, so they took measures that addressed this issue and improved the situation. Here we could distinguish two arguments: firstly, that disclosing gender gaps allows companies to analyze why the issue exists; secondly, it allows companies to solve the issue. However, as these two argumentsare so interlinked, it is easier to consider them as only one. Still, it is possible for argumentsin a paragraph to be very different, and so the paragraph will be annotated with more than one argument label. This is the case, for example, in the October 2017 text. There, the author speaks in favor of taking political action against unemployment. In paragraph four, they present their two main arguments: reducing unemployment reduces poverty and helps the overall economy of the country. These could be considered part of one single argument, as poverty levels can be considered a macroeconomic measure. However, the author seems to separate these two ideas, for example by mentioning how poverty affects children, thus making the poverty argument more linked to emotions, separating it from macroeconomic arguments.

The author presents their thesis at the beginning of the text: the thesis then needs to be accompanied by a description of the issue for the reader to understand it, and so the paragraph is annotated with the labels of those two functions. This is what happens, for example, in the October 2018 text. There, the author begins the text witha paragraph that essentially summarizes the whole text, saying that bats are dying out because of disease and accidents on wind farms (the issue), and that letting bats disappearharms agriculture and human health (the thesis).

Paragraph function summaries

In our annotation of the texts, aside from labelling each paragraph's function, we decidedto add a summary of the function. We can look at the December 2017 text for an example. There, the thesis (the second-hand clothes market is not as ethical as it might seem) can be found on paragraph three. Thus, before the thesis, there are two distinct paragraphs, but they perform the same function of introducing the issue being discussed in the text bysaying that thrifting or donating clothes seems like a "win-win", both cheap and ethical. As the two paragraphs are essentially transmitting the same idea, we can

summarize themboth as “The second-hand clothing market seems good for everyone, as it is an ethical choice and gives us cheap clothes”. After the thesis, the text contains seven other para- graphs. However, they can be summarized as three arguments: “The second-hand clothesmarket is more business than charity”, “The second-hand clothes market harms clothing manufacturers, especially in Africa” and “Donating clothes doesn’t end the problems of fast fashion, but it gives us a false idea that it does”.

We surmised that adding these summaries might be useful given the difference in length between the paragraphs and students’ potential answers. Perhaps comparing the answers against data of such a different length would not give the system enough information to evaluate the answers: due to the nature of the task, we decided to use a semantic similarity tool, and this tool was sensitive to text length. This precaution was later shown to be unnecessary, but it did help expedite the annotation process by helping the annotator recall the content of a paragraph quickly when referring to it to create answers.

Assessment criteria for student answers

We not only annotated the structure of SAT texts, butalso wrote possible answers that students might give if those texts were used for ST4. We wrote those answers to analyze how a dialogue system might evaluate them to return appropriate feedback. As we mentioned, creating the answers allowed us to control the variables that we wanted to analyze - the criteria for what type of feedback would be required for each answer. The assessment criteria for which we labelled the answers were: whether the students attempt to complete the task at all, whether the students use theirown words, whether the answer is complete, and whether it is correct. These are explained below.

On-task answers

The first of these criteria should require little explanation: before the system can evaluatehow the student is carrying out the task, it must detect whether they are indeed attempting to complete it. Only after the system knows that the student’s answer is actually an answercan it decide whether it is good or whether the student needs some corrective feedback. For example, if the student says “I need another explanation”, it would be useless to evaluate how this utterance summarizes the thesis and arguments of the reference text; the appropriate policy would be for the system to explain the task again.

Answers in the student’s own words

A good answer reflects that the student has understood the text, so that they will beable to use the information from the text for later stages in the task. A way of reflecting that one has truly understood some idea and is ready to evaluate it is to express it in one’s own words (Skidmore, in Mercer et al., 2019); therefore, one important aspect of a good answer is whether the student used their own words or copied the information in the text author’s voice. For example, if the text (in this example, the December 2016 text) says “The FAMILY Act is a commonsense measure whose time has come to modernizethe workplace to reflect the changing face of the American family, so that finally, we can relegate having to make the choice between earning a paycheck and caring for a loved oneto the dustbin of history where it belongs, a relic of a ‘Mad Men’ era gone by.”, and the student answers “The time has come for

the FAMILY Act to modernize the workplace to reflect the changing face of the American family, so that finally, we can relegate havingto make the choice between earning a paycheck and caring for a loved one to the dustbinof history where it belongs, a relic of a 'Mad Men' era gone by. ", the system needs to encourage them to analyze the ideas and put them into their own words.

Complete answers

ST4 is to serve as basis for the following stage in the task, where the student will have to argue their position in an issue; argumentation is also a key aspect of learning in dialogic pedagogy, as it is about exchanging views and evaluating ideas (Andrews, 2005), two processes that are key in dialogic teaching (ibid). For these reasons, it is important that students' answers are complete, meaning that the claims are backed by at least one premise, as giving reasons for a claim is both an indicator that dialogic teaching is taking place (Sedova, 2017) and a subskill involved in the development of oracy skills (Mercer et al., 2017). Backing claims with reasons is also a rule for discussions that students need to learn and reinforce for them to fully benefit from dialogic tasks in class (Mercer et al., 2010a). Continuing with the December 2016 text as an example, there the author defendsthe FAMILY Act, which would give workers paid leave for family reasons, claiming, among other arguments, that it is not fair for workers to have to choose between their family and their income, or that unpaid leave harms women disproportionately. Thus, if a student simply answers "The FAMILY Act is an important measure", the system will need to askthem to provide some justification.

Answers that summarize other elements (aside from the thesis and arguments), to provide more context for the claim and the premises, can also be considered correct, as long as they are not merely a lengthy summary of the whole text – if the student answersgiving all the information of the text, this might be a sign that they cannot distinguish the different elements of the text. Such lengthy answers have not been included; due to lack of time, other answer types have been prioritized which we believed more realistic.

Where there is no evident hierarchy among the arguments, we need to accept answersthat provide only one of those arguments; demanding more of the student would requirethem agreeing with the annotations in the dataset used by the dialogue system on how the argumentation of the text is structured. However, argument annotation is a complex taskwhere agreement is very difficult (Gurevych, 2021), so we need to keep the task simpleto avoid correcting the student on something caused by differences in subjectivity rather than an objective error (Kulatska, 2019). Moreover, a simple system would not be able to debate with the student whether other options might be correct (such capabilities would be ideal, but they are beyond the scope of our simplified system); therefore, the task and thefeedback need to be simplified. Only when the author favors one argument very visibly canthe system correct the student with some confidence for choosing a different, less relevant, argument. For example, if we look again at the December 2016 text, there the author uses several arguments to defend the FAMILY Act, giving workers paid leave, without apparently giving any of the arguments special attention over the others. Thus, an answerlike "The text defends a law that would allow people to get paid leave to take care of their family, so that people don't have to choose between their family and an income" could be considered perfect, even though it does not mention other arguments, such as the factthat other countries have paid leave for workers or that similar measures have not harmedbusinesses in the past.

An example of a text where a hierarchy of arguments is clearer is the March 2019 text, where the author claims that stricter safety regulations are needed for the transportation of oil by train. The author uses several arguments, but the one thatthey use in most paragraphs, one of them the first one, together with the thesis, is that not requiring oil companies to remove volatile gases before transportation can lead to deadlyaccidents. Therefore, if the student answers "Laws should force oil companies to remove gases from the oil that is transported by train. This process wouldn't be too difficult.", they would correctly be mentioning one of the author's arguments, that removing volatilegases is a simple process, but the system would need to help the student notice that thereis a more important argument that they have not mentioned.

Correct answers

Lastly, it is also important to remember that an argumentative text is not only a set of structural elements, but that these elements reflect a stance. For instance, in the annotation guidelines by Stab and Gurevych (2014), identifying the author's stance is seen as a crucialtask that is to be completed before analyzing the structure of the text, and it informs the identification of the different structural elements. The students' understanding of the author's stance is thus considered an important part of the educational task - their answerhas to not only deal with the same issues as the author's thesis in the text, but also discussthem in the same positive or negative light. Reflecting a different stance could be a sign that the student has misunderstood the text, so they might not be able to evaluate its ideasand bring them into discussion. For instance, we can look again at the March 2019 text, where the author says that oil companies need to be forced to remove volatile gases before transportation in order to avoid deadly accidents. There, if a student answered "The new safety regulations for transporting oil don't require companies to remove volatile gases. This was an acceptable decision - gases are not the main problem.", they would have completely misunderstood the text and would need some guidance.

Answer distribution and category labels

So that the dataset can serve to train the system to classify answers as good or bad according to our established assessment criteria, it would need to contain some perfect answers, as well as answers lacking in some of those criteria. A complete dataset could feature all possible combinations of criteria (i.e. perfect answers, answers lacking in one aspect, answers lacking in two, answers lacking in three, answers lacking in all four aspects). However, for a scaffolded task, feedback can focus on one aspect, so that the student can work on one skill at a time. For this reason, it is necessary to establish a hierarchy in the assessment criteria, so that the system can focus on the most important aspect when problems are detected concerning more than one assessment criteria. Naturally, the first step is determining whether the student's utterance is an answer. After that, we have established a hierarchy with completeness as the first aspect to be assessed, followed by paraphrasing, and ending with correctness:

1. **Completeness**
   The task preceding the summarization of the thesis and the premises involves spotting the paragraphs where that information is conveyed. Thus, completeness could be the first aspect to be assessed: if the student does not include all the

elements they spotted in the previous stage, that stage may need to be reviewed and the student reminded of the most important text chunks.

2. **Paraphrasing**
   Once the student knows what information they need to convey, we would need to assess whether they understood the ideas contained in the paragraphs they pointed to: this would involve both paraphrasing and accurately conveying the author's stance. As putting other's ideas into one's voice can be seen as a vehicle towards incorporating those ideas into our understanding (Skidmore, in Mercer et al., 2019), we could consider the paraphrasing aspect earlier in the hierarchy. This would leave the aspectof correctness third and last in the hierarchy.

3. **Correctness**
   A student who has understood which bits of information they need to use and whohas put that information into their own words could be expected to only rarely make any errors in interpreting the stance conveyed in what they just read. Still, this possibility has to be contemplated. Such cases might reveal that the student's reading comprehension is not at the level of the task, and perhaps it might be useful to add a multiple-choice subtask that brought the student's attention to specific points in the text and helped them see which stance they convey. Opting for the multiple-choice format with no Natural Language Understanding could make the integration of the subtask easier (Jokinen, 2009), though it would require designingthe questions and possible answers for each text used by the dialogue system. Perhaps a good aspect-based-sentiment analysis tool could help automate this task, as the tool could recognize which expressions conveyed a positive or negative stance and bringthe student's attention to those specific parts of the text. That, however, is to be analyzed in future studies.

With this hierarchical approach in mind, as well as to also ease our dataset compilation efforts, we have only fabricated perfect answers and answers lacking in one aspect (i.e. answers that are off-task, answers that are incomplete, answers that are incorrect, and answers that are copying the text instead of paraphrasing it). We acknowledge the limitations of this approach: though the system is meant to focus on one assessment aspect at a time, real student answers may present issues with more than one assessment criteria.Detecting the issue highest in the hierarchy in such answers that present more than one problem might complicate the classification task for the system: e.g. if an answer is a poorparaphrase will the system also detect that it is incomplete and apply the policy for returning feedback concerning completeness, the higher aspect in the hierarchy? Nonetheless, we believe it sensible to start our tests with simpler data and only test how more complexanswers are classified once we confirm whether simple answers can be accurately classified.

We also need to establish the criteria for the design of the answers, beyond the typeof error they may feature, as that is too vague a guideline - the dataset could become very imbalanced and it might be hard to extract clear conclusions from an experiment. Below we list the labels for each answer type and explain how answers with that label were created:

- Off-task questions (**offt**): these are questions that do not attempt to complete thetask, either because the student is confused or because they do not want to do thetask.

- **offt-needhelp-questask**: Direct question about task instructions (e.g. "What doI need to do now?").
- **offt- needhelp-commtask**: Command to clarify something (e.g. "Gimme some help").
- **offt- needhelp-conf**: Expression of confusion (e.g. "Eh?").
- **offt-gibb**: Random utterance, here in the form of random song lyrics (e.g. "Heart beats fast Colors and promises How to be brave? How can I love when I'm afraid to fall?").
- **offt-rand**: Random/off-task utterance, but with at least one word from the domain of the reference text (e.g in the December text about the FAMILY Act, "Never heard of this FAMILY Act").
- **offt-disg**: Expression of disgust at task (e.g. "This suuuckss").
- **offt-pers**: Question about the system's personality (e.g. "Are you an evil robot?"). Bii et al. (2013) observed that these occurred with some frequency among high-school students using a chatbot.

- Perfect answers (**pa**): answers that are complete (state the author's thesis and at least one way it is justified), good paraphrases (they do not copy the text, nor dothey simply replace or move a few words), and that are correct (state the author's position on the issue, and not an opposed view). For example, for the December 2017 text about how the second-hand clothes industry is not as ethical as it seems,
  "Most people believe that donating clothes is a good thing, but they're harming some businesses and the planet cos it's just putting a patch on fast fashion, not solvinganything" would be a good answer.

- Poor paraphrases (**poorpar**): answers that take a text chunk containing the thesis and simply move some words or replace them with synonyms to make it less of anexact copy. When the thesis is stated in more than one paragraph, we take them allas reference for an equal number of answers (e.g. if the thesis is in two paragraphsand a dataset category is meant to have four answers, half will poorly paraphrase one paragraph, and the other two answers will poorly paraphrase the other paragraph).The poor paraphrases that used synonyms were created using Wordnet to expedite the process, despite this resulting in strange utterances (still, replacing words with inadequate synonyms is something that students might in fact do, if perhaps not tothe same extent).

  - **poorpar-syn2**: Replacing 20% of words in the paragraph with synonyms. A short example that we can show here comes from the May 2017 text, where the thesis can be found on paragraphs eight and twelve. Paragraph twelve says "With so much to gain, we need to cut work hours while there is still time.",and with 19% of the words replaced it becomes "With therefore much to profit,we motivation to cut work hours while there is still time".
  - **poorpar-syn5**: Replacing 50% of words in the paragraph with synonyms. The previous example paragraph would now become "With then much to addition,we necessitate to hack work hour while there be still time" - it was not

always possible to find synonyms for half the words, especially in short paragraphs.

  - **poorpar-ord1**: Making a small change in the paragraph's structure. Continuing with the May 2017 example, paragraph twelve becomes "With so much to gain,we need to cut work hours", where we merely removed the end of the sentence.
  - **poorpar-ordmore**: Making several changes in the paragraph's structure. For example, here paragraph twelve becomes "We need to cut work hours. There is much to gain, but we need to make sure there is still time.", where we reordered the sentence and changed the connectors between the sentence parts.

- Incomplete answers (**com**): Answers that are missing the text thesis or its justification, or both (but are genuine attempts at completing the task). To illustrate whattypes of answers are assigned this label, we will be using examples from the October 2017 text, where the author defends taking political action against unemployment to reduce poverty and improve the economy.

  - **com-nojus**: Stating the author's stance without justification (e.g. "Congress needs to pass a bill that reduces unemployment").
  - **com-offfocus**: Summarizing text parts that merely provide context (e.g). "FDR signed the so-called second bill of rights to reduce unemployment".
  - **com-minarg**: Stating the author's thesis but justifying it with a minor argument. This category only applies to texts with a clear hierarchy of arguments. An example would be "Policy makers need to tackle unemployment. That could also solve infrastructure issues, killing two birds with one stone".
  - **com-arg**: Stating an argument without linking it to the thesis (e.g. "If people have no job they can't feed their family").
  - **com-topic**: Describing the topic without stating a position (e.g. "The text discusses the consequences of unemployment and what Congress can do about it").

- Incorrect answers (**corr**): Answers that contain a thesis and justify it, but these elements contradict the author's position. For this category we'll be showing examplesfrom the October 2018 text, where the author claims that we should protect bats from disease and accidents, because bats provide excellent pest control for agriculture and disease control.

  - **corr-negthes**: Inverting the author's stance (e.g.) "We shouldn't put our efforts into protecting bats, because we have other means of killing insects.".
  - **corr-negarg**: Stating the author's thesis but negating the justification (e.g. "Pesticides are safer and more effective than bats for getting rid of pests in agriculture, but we still gotta preserve them as any other endangered species.").

– **corr-negall**: Inverting the author's thesis and its justification (e.g. "It appears that bats are dying out, but humans are in no position to solve this. Also, bats are a very harmful species, so it's no big loss.").

The distribution of answer types is shown in Figure 2. The first decision was which proportion of answers would need to be on- and off-task. Preliminary analyses suggested that distinguishing between on- and off-task answers could be easy, and thus fewer off-task answers might be needed. Inside the category of off-task answers, all sub-categories were given a similar weight, giving only slightly more weight to the off-rand category (whereoff-task answers contain a domain word), to have sufficient data to draw conclusions on the system's reliance on single appearances of domain words. On-task answers include both perfect answers and answers with one problematic aspect. Inside each of the three categories of aspects that we considered for feedback there were many subcategories; therefore, to have enough data to draw conclusions on each of those sub-categories, each of the three categories had to be assigned a sizeable proportion of the dataset. This left the category of perfect answers with only 10% of the total; however, as it has no subcategories, this might be sufficient. The completeness category is the one of the three feedback categories with the most subcategories, accounting for the main ways in which a student may fail to include a thesis and a premise to back it in their answer. For this reason, as well as thefact that completeness has been assigned first place in the hierarchy of feedback types, this category has been assigned the largest proportion of the dataset. The proportions assignedto the other two feedback type categories respond to the same reasoning. Firstly, whereas completeness has five subcategories, paraphrasing has four, and correctness three. Also,in the hierarchy of feedback types, completeness was put in first place, paraphrasing insecond, and correctness in third.

We must acknowledge that this category distribution is not perfectly balanced; some categories have been assigned a larger proportion of the answers when we believed that classifying those categories would be more challenging. To be able to create a number of answers that could make the dataset usable (i.e. allowing us to extract some initial conclusions, despite how limited they might be), we needed to simplify the creation process. Real student answers could be expected to belong to the negative class for more than one of our classification criteria (on-/off-task, completeness, correctness, paraphrasing). However, creating more realistic answers that failed in more than one aspect while maintaining abalanced distribution would have slowed down the fabrication of answers and reduced output significantly (e.g. creating incomplete answers that only contained the thesis and no argument was easy, but having to make some of them be also bad paraphrases while keeping count how many answers need to be in each class would have required much more time or more annotators). Nonetheless, as will be discussed in section 6.2.3, the category distribution does not seem to have a strong impact on our tests.

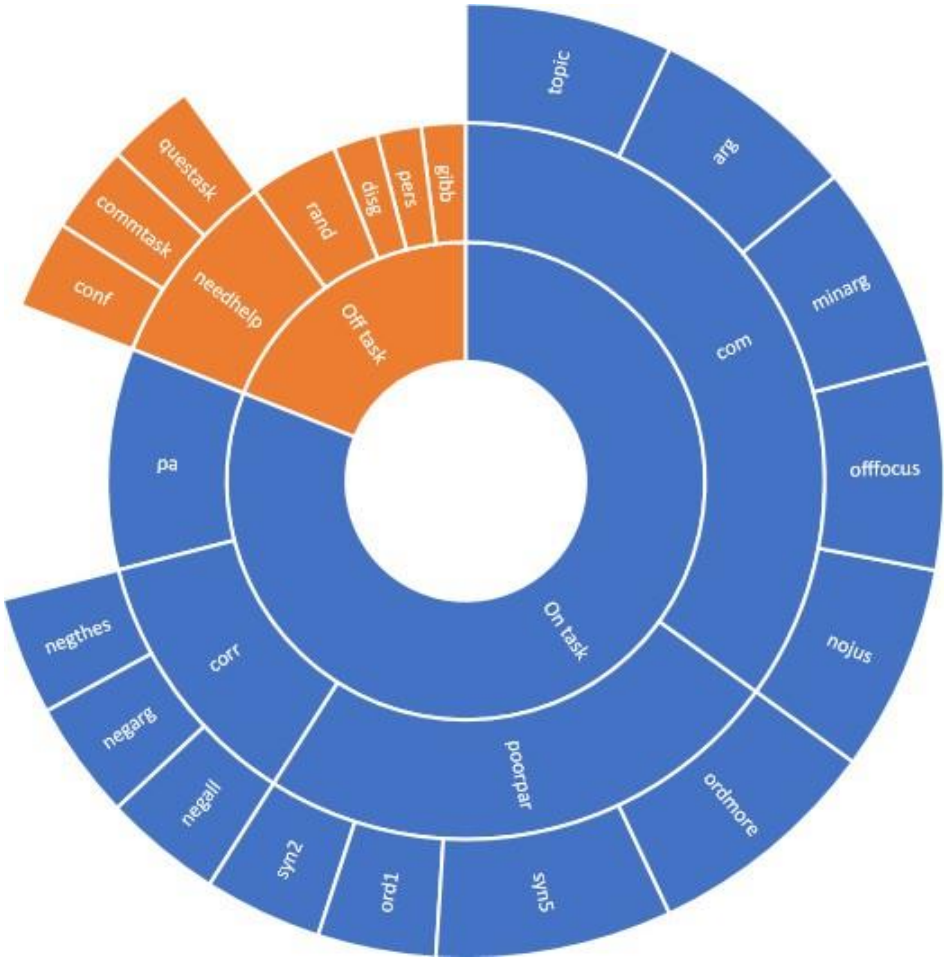

Figure 2: Category distribution

Dataset size

The size of the dataset was decided merely based on what was feasible in the time available. In the end, we were able to annotate ten SAT Essay texts for their argumentative structure.For each text, 93 or 100 answers were created - one answer type was not applicable to all texts, as explained in the previous section, resulting in six texts having 93 answers instead of 100. We acknowledge the limited size of the dataset, but this was enough to reach some initial conclusions from an experiment.